UTBNLP at CLEF JOKER 2025 Task 2: mBART-50 Fine-Tuning with Dictionary-Guided Forced Decoding and Phoneme-Based Techniques for English-French Pun Translation*

Notebook for the UTBNLP Lab at CLEF 2025

Duvan Andres Marrugo-Tobon[1,*], Jeison D. Jimenez[1], Jairo E. Serrano[1], Juan C. Martinez-Santos[1] and Edwin Puertas[1]

[1]Universidad Tecnológica de Bolívar, School of Digital Transformation, Cartagena de Indias 130010, Colombia

Abstract

This paper presents a pun-focused translation system developed for the JOKER-2025 Task 2 competition on English-to-French pun translation. The system combines mBART-50 fine-tuning, LoRA adapter training, and a novel phoneme-aware forced-decoding strategy guided by a specialized pun dictionary. The end-to-end pipeline encompasses robust pun detection and tagging, oversampling-based data augmentation, phoneme transcription, and sentiment feature enrichment, enabling comprehensive capture of the layered meanings and playful ambiguity inherent in puns. Multiple model configurations and decoding schemes were evaluated, with the best configuration achieving a BLEU score of 32.82 and ranking 22nd overall. These findings underscore the effectiveness of incorporating phonetic cues and lexical constraints to enhance both faithfulness and humor preservation in pun translation across languages.

Keywords

Pun translation, mBART-50, phoneme transcription, sentiment analysis.

1. Introduction

The automatic translation of humor remains one of the most persistent challenges in natural language processing (NLP), especially when it comes to wordplay such as puns. Recent studies [1] [2] highlight that humorous content demands systems capable of navigating a delicate balance between semantic fidelity and creative linguistic adaptation. Unlike general text, humor heavily relies on implicit cultural knowledge, phonetic ambiguity, and contextual cues that do not always transfer directly across languages.

Among humorous forms, puns stand out for their intricate layering of meanings through phonological resemblance, lexical ambiguity, or syntactic manipulation. This dual (or multiple) reading effect often exploits language-specific structures, making literal translation ineffective or outright impossible [3]. The translation of puns thus poses challenges not only for automatic systems but also for human translators, requiring a blend of linguistic creativity and cultural sensitivity.

In response to these challenges, shared tasks like JOKER [4, 5, 6] have provided a standardized benchmark for the computational humor community. They promote the development of new datasets, evaluation metrics, and innovative translation strategies specifically tailored to preserve wordplay across languages. These tasks have revealed that traditional neural machine translation (NMT) models, optimized for literal semantic transfer, tend to underperform when the goal shifts from factual

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*This paper was prepared for participation in the JOKER 2025 shared task on wordplay translation at CLEF 2025.

*Corresponding author.

marrugod@utb.edu.co (D. A. Marrugo-Tobon); jalvear@utb.edu.co (J. D. Jimenez); jserrano@utb.edu.co (J. E. Serrano); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

0009-0002-9537-0209 (D. A. Marrugo-Tobon); 0009-0001-0134-8426 (J. D. Jimenez); 0000-0001-8165-7343 (J. E. Serrano); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

correctness to stylistic and pragmatic equivalence.

Recent developments in multilingual sequence-to-sequence modeling, most notably the mBART-50 architecture combined with parameter-efficient adaptation techniques such as LoRA adapters [7, 8] have unlocked powerful domain-specific translation capabilities. Complementing these modeling advances, work on lexically constrained decoding [9] and full phoneme-level pipelines [10] has demonstrated that explicit control over the generation process can help retain critical structural elements in challenging tasks, such as pun translation.

A variety of specialized methods for handling wordplay have since emerged:

- Dual-attention architectures that localize and interpret puns via gloss pairs [11].
- GPT-3–based generators exploiting contextual ambiguity for creative pun synthesis [12].
- Enriched corpora annotated with keywords and explanatory notes to boost classification and generation quality [13].
- Unified frameworks addressing both homophonic and homographic puns [14].
- Grapheme-to-phoneme augmentation techniques preserving acoustic wordplay signals [15].
- Large-model evaluations exposing "lazy" pun reproduction and motivating new metrics [16].

While these approaches advance semantic disambiguation, data balancing, and lexical enforcement, few integrate phonetic features directly into the decoding mechanism itself.

This paper proposes an end-to-end pipeline for English-to-French pun translation that unites mBART-50 fine-tuning (with optional LoRA adapters), systematic pun detection and markup, explicit G2P-based phoneme transcription, and a phoneme-aware forced-decoding strategy guided by a curated pun dictionary. By enriching inputs with sentiment and phonetic embeddings, experiments show substantial gains in preserving both the humorous intent and the phonological wordplay of source puns. This paper is structured as follows: Section 2 describes the methodology, including dataset description, pun detection and tagging, data augmentation, feature enrichment, and model training. Section 3 presents performance results and qualitative analysis of the pipeline variants. Finally, Section 4 offers conclusions and outlines directions for future work.

2. Methodology

This section describes the architecture of the English-to-French pun translation system developed for Task 2 of the JOKER-2025 competition. As shown in Figure 1, the workflow comprises five main stages:

- **Data Preprocessing:** Automated detection of puns in the source text and insertion of explicit markup to guide translation.
- **Feature Engineering:** Enrichment of input data with sentiment labels and phonemic transcriptions to capture both emotional tone and sound-based wordplay.
- **Model Training:** Comparison of two complete fine-tuning training strategies versus the LoRA adaptation that is efficient in parameterizing applied to the multilingual model mBART-50.
- **Inference:** Application of a custom forced decoding algorithm, guided by a pun-specific lexicon, to ensure the preservation of key wordplay elements during translation.
- **Evaluation:** Performance assessment using both automatic metrics and manual evaluation criteria defined by the JOKER track, with an emphasis on maintaining the intended humorous effect.

2.1. Dataset Description

The dataset used in this work was released as part of the JOKER 2025 shared task on wordplay translation [6]. It consists of English-French translation pairs curated to evaluate pun translation quality, with

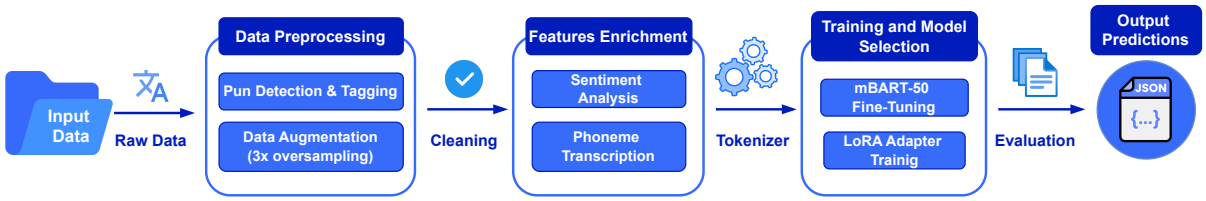

Figure 1: System pipeline for English-to-French pun translation.

each entry formatted as a JSON object containing: `id_en` (unique identifier), `en` (source English text containing wordplay), and `fr` (French translation).

A distinctive characteristic of the training data is the presence of multiple French translations for each English pun, reflecting different approaches to wordplay preservation. The dataset contains training examples similar to those provided in previous editions, offering diverse examples of pun translation strategies.

The test dataset follows the same format but contains only the original English text (`en`) and identifier (`id_en`), requiring systems to generate French translations that preserve both semantic meaning and humorous effect. The evaluation prioritizes wordplay preservation over traditional translation metrics, following Delabastita's pun translation strategy. Table 1 shows the composition of the dataset provided by the competition organizers.

Table 1
JOKER-2025 Task 2 Dataset Composition

Split	Instances	Format
Training	5838	JSON with multiple FR translations
Test	4537	JSON with EN text only

2.2. Data Preprocessing

The preprocessing pipeline implements a systematic approach to prepare the pun translation dataset for model training. As illustrated in Figure 1, the preprocessing stage consists of two sequential operations designed to enhance wordplay structure representation and address dataset imbalances inherent in pun translation tasks.

Pun Detection and Tagging

English puns are automatically identified by matching the input text against a curated lexicon, and each matched occurrence is wrapped with special delimiters to guide the translator, as shown in Algorithm 1.

Key points

- Each input sentence is iterated over and scanned against the English to French pun dictionary.
- On the first case-insensitive match, <PUN_EN> is injected before and after the English pun.
- During training the French side can receive <PUN_FR> around the mapped French entry.
- This explicit markup focuses the downstream model on the wordplay span without disturbing the surrounding context.

Algorithm 1 Pun Detection and Tagging

Require: List of sentences $\{s_i\}$, pun lexicon $\mathcal{L} = \{(e, f)\}$ where e is English pun, f its French equivalent

Ensure: Tagged sentences with ⟨PUN_EN⟩, ⟨PUN_FR⟩

for each sentence s in input **do**
  `tagged` $\leftarrow s$
  for each (e, f) in $\mathcal{L}$ **do**
    if lower(e) occurs in lower(s) **then**
      Mark first match: `tagged` $\leftarrow$ `replaceOnce(` `tagged`, e, ⟨PUN_EN⟩ e ⟨PUN_EN⟩`)`
      break ▷ Only tag one pun per sentence
    end if
  end for
  Output `tagged`
end for

Tagging Example

Original:
`A skier retired because he was going downhill.`

Tagged:
`A skier retired because he was going <PUN_EN> downhill <PUN_EN>.`

Data Augmentation

Targeted oversampling is applied to mitigate the imbalance between punny and non-punny examples. Each pun-containing instance is replicated three times in the training set. This ensures a balanced representation of wordplay cases during learning, following established minority-class techniques to improve the model's ability to handle scarce patterns.

As illustrated in Figure 2, the augmentation process operates systematically through four sequential stages.

- **Pun Identification:** Identify all examples containing at least one detected pun.
- **Data Augmentation:** Create three additional copies of each example containing puns.
- **Dataset Consolidation**: Combine original and augmented examples into a unified training set.
- **Dataset Shuffling**: Shuffle the dataset using a fixed random seed for reproducibility.

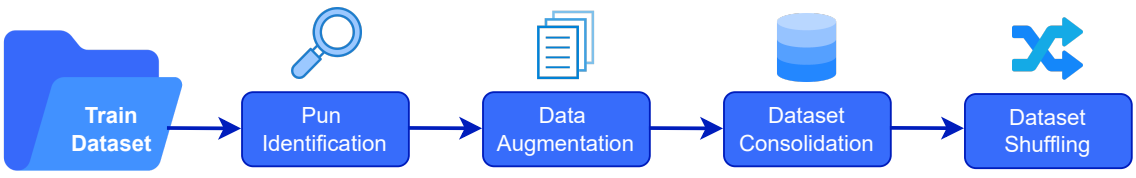

Figure 2: Overview of the data augmentation strategy for addressing class imbalance in pun translation training.

The augmentation process creates a more balanced training distribution, allowing the fine-tuned model to develop robust representations for both standard translation patterns and specialized wordplay mechanisms. Without this rebalancing, the model would predominantly optimize for literal translation patterns, potentially failing to preserve wordplay in test scenarios.

2.3. Features Enrichment

A comprehensive feature engineering approach is implemented to incorporate linguistic characteristics relevant to wordplay translation. This stage enriches the preprocessed dataset with semantic and phonetic features designed to capture emotional and acoustic patterns that may influence preservation strategies for puns.

2.3.1. Sentiment Analysis

Each training instance is augmented with sentiment classification to capture the emotional context of humorous content. Sentiment analysis is performed by a pre-trained transformer model (cardiffnlp/twitter-roberta-base-sentiment-latest) fine-tuned on social media data [17], providing robust performance for informal and creative text typical in pun datasets.

The sentiment classifier assigns one of three labels to each original English sentence: POSITIVE, NEGATIVE, or NEUTRAL. This categorical sentiment information is stored in a new field in the dataset, providing the translation model with explicit emotional context that can guide the preservation of the appropriate tone in the target language.

2.3.2. Phonemic Transcription

Automatic phonemic transcription is implemented using a grapheme-to-phoneme (G2P) conversion system [15] to capture phonetic patterns relevant to sound-based wordplay. The G2P model converts English text to ARPAbet phonetic notation, providing explicit acoustic representations that may inform the model about sound-based pun mechanisms. The phonemic transcription is stored in a new field, enabling the model to access pronunciation patterns that are often crucial for wordplay preservation, particularly for puns based on homophones or rhymes. Each training record contains a unique identifier, the English pun, and one or more human-verified French translations. For feature enrichment, each sentence is paired with its phonemic transcription in ARPAbet notation to highlight sound-based wordplay structures. Algorithm 2 summarizes the phoneme extraction process.

Algorithm 2 Detailed Feature Enrichment Process

Require: Dataset of pun sentences with unique IDs and English text
Ensure: The same dataset with added sentiment and phoneme annotations

1: **for** each record in the dataset **do**
2: Extract the English text
3: Preprocess the text:
- Lowercase normalization
- Removal of extra whitespace or unwanted characters

4: Predict the sentiment label (POSITIVE, NEGATIVE, NEUTRAL) using a pre-trained sentiment classification model
5: Generate the phoneme transcription:
- Tokenize the text into words
- For each word, obtain its ARPAbet phoneme sequence
- Reconstruct the full sentence phoneme string, preserving punctuation

6: Add the sentiment label and phoneme string to the record
7: Leave the translation field empty for future prediction
8: **end for**
9: **return** Enriched dataset in JSON format with fields: `id_en`, `en`, `sent`, `phon`, `fr`

Each input pun undergoes careful preprocessing, sentiment classification, and phoneme conversion to enrich its representation for training. This step enables the translation model to capture the emotional

tone and sound-based ambiguities inherent in puns. Algorithm 2 describes this process in detail, and an example output is presented below.

Example of a Fully Enriched Record

```
 id_en: en_1
en: Save the whales, spouted Tom.
sent: POSITIVE
phon: S EY1 V  DH AH0  W EY1 L Z ,  S P AW1 T AH0 D  T AA1 M .
fr: Sauvez les baleines, souffla Tom.
```

2.4. Model Training

Two distinct fine-tuning strategies for pun translation are evaluated, comparing complete parameter optimization against parameter-efficient adaptation techniques. Both approaches utilize mBART-50 (facebook/mbart-large-50-many-to-many-mmt) as the base multilingual transformer [18], configured for English-to-French translation with source language token en_XX and target language token fr_XX.

Algorithm 3 details the training procedure for both full fine-tuning and LoRA adaptation.

Algorithm 3 Model Training for Pun Translation

Require: Enriched training data, pre-trained mBART-50 model
Ensure: the Fine-tuned model is ready for pun translation

```
Initialize the mBART-50 model and tokenizer
Define the objective: maximize translation accuracy while preserving wordplay
Tokenize sentences and truncate to maximum length
Create mini-batches and set optimizer and learning rate scheduler
if Full fine-tuning then
    Update all model weights
else
    Freeze base weights and train only low-rank adapters (LoRA)
end if
for each training epoch do
    Compute loss on each batch
    Backpropagate gradients and update weights
    Periodically evaluate the validation set
end for
Select best checkpoint based on validation performance
return Trained model for inference
```

2.4.1. Full Parameter Fine-tuning

The first approach implements complete model fine-tuning, updating all parameters of the mBART-50 architecture during training. This method provides maximum model expressivity by allowing unrestricted parameter updates across all transformer layers, attention mechanisms, and feed-forward networks.

The training process employs the Seq2SeqTrainer framework with the hyperparameter configuration shown in Table 2.

The complete fine-tuning approach enables comprehensive adaptation to the pun translation domain, allowing the model to learn specialized linguistic patterns for wordplay preservation across all network parameters. This method provides maximum learning capacity by updating the entire transformer architecture.

Table 2
Training configuration parameters for fine-tuning

Parameter	Value
Batch size	8
Training epochs	10
Precision	FP16
Optimizer	AdamW
Maximum sequence length	128

2.5. LoRA Adapter Training

The second approach implements Low-Rank Adaptation (LoRA) [7], a parameter-efficient fine-tuning technique that freezes the pre-trained model weights and trains only low-rank decomposition matrices inserted into attention layers. This method significantly reduces the number of trainable parameters while maintaining competitive performance.

The LoRA configuration targets attention projection matrices with the settings detailed in Table 3.

Table 3
LoRA Configuration Parameters

Parameter	Value
Rank (r)	8
Alpha (α)	32
Dropout	0.1
Target modules	q_proj, v_proj, k_proj, o_proj
Task type	SEQ_2_SEQ_LM

The adapter training maintains the same core hyperparameters as full fine-tuning but operates on a drastically reduced parameter space. The training configuration turns off intermediate checkpointing and external logging for efficiency, thereby concentrating computational resources on the adapter optimization process.

Both training strategies incorporate the enhanced dataset with pun tagging and linguistic features, enabling comparison between maximum parameter flexibility and efficient domain adaptation for specialized translation tasks. The following boxed example illustrates how the trained model processes an enriched input step by step. It demonstrates how tokenization, phoneme cues, sentiment guidance, and forced decoding work together to generate a pun-preserving French translation.

Step-by-Step Example: From Enriched Input to Final Translation

Input Record:

```
id_en: en_1
en: Save the whales, spouted Tom.
sent: POSITIVE
phon: S EY1 V  DH AH0  W EY1 L Z ,  S P AW1 T AH0 D  T AA1 M .
```

Model Process:

1. **Tokenization:** The English sentence is tokenized with special language tokens (e.g., `en_XX`, `fr_XX`).
2. **Encoding:** The tokenized input is converted into embeddings using mBART-50's encoder layers.
3. **Forced Decoding:** During generation, the decoder applies lexical constraints guided by the pun dictionary and phoneme cues.
4. **Sentiment Guidance:** The model conditions the output tone to match the predicted sentiment (here: POSITIVE).
5. **Output:** The decoder generates a French pun translation that preserves the wordplay.

Final Translation:

```
fr: Sauvez les baleines, souffla Tom.
```

3. Performance Results

The JOKER-2025 Task 2 competition for English-to-French pun translation was addressed using a progressive pipeline comprising three configurations: a fully fine-tuned system and two parameter-efficient LoRA variants. Table 4 summarizes the official BLEU scores assigned during the automatic evaluation phase.

Table 4
Official JOKER-2025 Task 2 scores for our pipeline variants.

Rank	Approach	Score (BLEU)
22	Full Fine-tuning	32.82
27	Baseline LoRA	28.17
32	Refined LoRA	27.57

The results indicate that the *Full Fine-tuning* model substantially outperformed both LoRA-based variants, achieving approximately 4-5 points higher BLEU. It suggests that, for pun translation, extensive parameter updates allow the model to learn better the subtle interplay of lexical form and multiple senses typical of wordplay.

However, as established in the JOKER evaluation framework, BLEU alone cannot reliably measure the success of pun translation, which requires preserving both semantic coherence and creative linguistic distortions. Therefore, all submitted systems are additionally evaluated by human experts using criteria including lexical field preservation, pun form retention, appropriateness of style shift, humor maintenance, and the detection of syntactic or lexical inaccuracies. Ultimately, the final ranking prioritizes the number of translations that successfully transfer the original wordplay mechanisms.

Example of Successful Pun Preservation (Full Fine-tuning)

English: "A skier retired because he was going downhill."
Phonemes: `AH0 S K IY1 ER0 R IH0 T AY1 ER0 D B IH0 K AO1 Z IY1 W AA1 Z G OW1 IH0 NG D AW1 N HH IH1 L`
French (Full Fine-tuning): "Le skieur a pris sa retraite car il n'arrivait plus à remonter la pente."
Analysis: The model correctly preserves the double sense of "going downhill" (literal skiing + figurative decline) by using the idiomatic French expression "remonter la pente", demonstrating strong handling of layered meaning and phonetic nuance.

Example of Failed Pun Preservation (Baseline LoRA)

English: "Someone once accused me of stealing an old, rare, valuable stamp, and I philately denied it."
Phonemes: `S AH1 M W AH1 N AH1 N S AH0 K Y UW1 Z D M IY1 AH0 V S T IY1 L IH0 NG AE1 N OW1 L D R EH1 R V AE1 L Y UW0 B AH0 L S T AE1 M P AE1 N D AY1 F IH0 L AE1 T L IY0 D IH0 N AY1 D IH0 T`
French (Baseline LoRA): "Quelqu'un m'a accusé d'avoir volé un timbre ancien et rare, et je l'ai nié sans hésiter."
Analysis: Here, the pun on "philately" (which plays on "flatly denied") is entirely lost. The baseline LoRA output conveys a literal denial, missing both the wordplay and the sound-based twist. It shows LoRA's limits without explicit phoneme constraints.

Combining automatic and qualitative results, the team achieved 5th place overall, with the highest BLEU of 32.82 for the Full Fine-tuning system. It confirms that extensive parameter updates yield more robust pun preservation. In contrast, the Baseline and Refined LoRA models, despite being computationally lighter, often fall back on literal phrasing, primarily when the pun hinges on phonetic similarity or double meanings. The contrast demonstrates that creative phenomena, such as puns, demand models to learn deeper associations across syntax, semantics, and phonology, which lightweight adapters alone may fail to capture fully.

Therefore, while BLEU provides an initial signal, manual expert evaluation remains crucial for fair ranking, as it verifies whether both the form and sense of the original wordplay survive translation. These insights support the adoption of hybrid approaches, which combine efficient adaptation with advanced decoding and post-editing to bridge the gap between automatic translation and human-like linguistic creativity.

4. Conclusions and Future Work

Our system *Full Fine-tuning* achieved the highest BLEU (32.82, 22nd place) by fully updating the mBART-50 weights, which allowed it to capture both the semantic and phonetic layers of puns (e.g. 'going downhill'*remonter la pente*). However, this approach incurs significant training and inference costs and still occasionally falters in out-of-vocabulary or heavily constrained puns. In contrast, the two LoRA variants (BLEU 28.17 and 27.57) required far fewer resources but consistently defaulted to literal renderings when phonetic or lexical cues were subtle (e.g., 'philately denied'plain denial), exposing their limited capacity to model creative wordplay.

These results underscore two key lessons: First, BLEU, while useful, only partially reflects the success of humor translation, as high overlap does not guarantee the preservation of wordplay form or tone shift. Second, manual expert evaluation remains indispensable to assess lexical field preservation, stylistic

shifts, and the maintenance of humorous effect. Only human judgments can reliably rank systems by their ability to transfer both the form and the sense of puns across languages.

Future directions include:

- **Dynamic Phoneme Constraints**: integrate on-the-fly phoneme matching to handle unseen puns.
- **Pun-Aware Metrics**: develop semi-automatic measures that correlate with human judgments of wordplay.
- **Adapter Enhancements**: augment lightweight LoRA with explicit phoneme–grapheme alignment layers to narrow the quality gap.
- **Cross-Domain Extension**: apply and evaluate the pipeline on other language pairs and humor forms (idioms, jokes).

Acknowledgments

The authors would like to acknowledge the support provided by the master's degree scholarship program in engineering at the Universidad Tecnologica de Bolivar (UTB) in Cartagena, Colombia.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. A. Aamir, C. C. Moreno, S. Sundelin, J. Biznárová, M. Scigliuzzo, K. E. Patel, A. Osman, D. Lozano, I. Strandberg, S. Gasparinetti, Engineering symmetry-selective couplings of a superconducting artificial molecule to microwave waveguides, Physical Review Letters 129 (2022). URL: http://dx.doi.org/10.1103/PhysRevLett.129.123604. doi:10.1103/physrevlett.129.123604.

[2] T. Hossain, S. Dev, S. Singh, MISGENDERED: Limits of large language models in understanding pronouns, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5352–5367. URL: https://aclanthology.org/2023.acl-long.293/. doi:10.18653/v1/2023.acl-long.293.

[3] B. Khalid, M. Alikhani, M. Stone, Combining cognitive modeling and reinforcement learning for clarification in dialogue, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 4417–4428. URL: https://aclanthology.org/2020.coling-main.391/. doi:10.18653/v1/2020.coling-main.391.

[4] L. Ermakova, R. Campos, A. Bosser, T. Miller, Overview of the CLEF 2025 JOKER lab: Humour in machine, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[5] L. Ermakova, et al., Overview of the CLEF 2025 JOKER task 1: Humour-aware information retrieval, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[6] L. Ermakova, et al., Overview of the CLEF 2025 JOKER task 2: Wordplay translation from english into french, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[9] S. Braun, O. Vasilyev, N. Iskender, J. Bohannon, Does summary evaluation survive translation to other languages?, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2425–2435. URL: https://aclanthology.org/2022.naacl-main.173/. doi:10.18653/v1/2022.naacl-main.173.

[10] X. Zhao, E. Durmus, D.-Y. Yeung, Towards reference-free text simplification evaluation with a BERT Siamese network architecture, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13250–13264. URL: https://aclanthology.org/2023.findings-acl.838/. doi:10.18653/v1/2023.findings-acl.838.

[11] S. Liu, M. Ma, H. Yuan, J. Zhu, Y. Wu, M. Al-Ajarmah, A dual-attention neural network for pun location and using pun-gloss pairs for interpretation, 2021. doi:10.48550/arXiv.2110.07209.

[12] A. Mittal, Y. Tian, N. Peng, AmbiPun: Generating humorous puns with ambiguous context (2022) 1053–1062. URL: https://aclanthology.org/2022.naacl-main.77/. doi:10.18653/v1/2022.naacl-main.77.

[13] J. Sun, A. Narayan-Chen, S. Oraby, A. Cervone, T. Chung, J. Huang, Y. Liu, N. Peng, ExPUNations: Augmenting puns with keywords and explanations (2022) 4590–4605. URL: https://aclanthology.org/2022.emnlp-main.304/. doi:10.18653/v1/2022.emnlp-main.304.

[14] Y. Tian, D. Sheth, N. Peng, A unified framework for pun generation with humor principles (2022) 3253–3261. URL: https://aclanthology.org/2022.findings-emnlp.237/. doi:10.18653/v1/2022.findings-emnlp.237.

[15] A. Pine, P. William Littell, E. Joanis, D. Huggins-Daines, C. Cox, F. Davis, E. Antonio Santos, S. Srikanth, D. Torkornoo, S. Yu, G_i2P_i rule-based, index-preserving grapheme-to-phoneme transformations, in: S. Moeller, A. Anastasopoulos, A. Arppe, A. Chaudhary, A. Harrigan, J. Holden, J. Lachler, A. Palmer, S. Rijhwani, L. Schwartz (Eds.), Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 52–60. URL: https://aclanthology.org/2022.computel-1.7/. doi:10.18653/v1/2022.computel-1.7.

[16] Z. Xu, S. Yuan, L. Chen, D. Yang, "a good pun is its own reword": Can large language models understand puns?, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 11766–11782. URL: https://aclanthology.org/2024.emnlp-main.657/. doi:10.18653/v1/2024.emnlp-main.657.

[17] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, J. Camacho-collados, TimeLMs: Diachronic language models from Twitter, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 251–260. URL: https://aclanthology.org/2022.acl-demo.25. doi:10.18653/v1/2022.acl-demo.25.

[18] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan, Multilingual translation with extensible multilingual pretraining and finetuning, arXiv preprint arXiv:2008.00401 (2020).